**Fundamentals of Multimedia Computing**

Chapter 2: Elements of Multimedia Computing

**Authors:**

**Gerald Friedland**

**and**

**Ramesh Jain**

**Draft for Comments**

**Chapter 2**

**Elements of Multimedia Computing**

**Introduction**

~~Multimedia is closely related to how humans experience the world.~~ **Multimedia reflects how humans experience the world.** In this chapter, we introduce how people interpret and use sensory signals to understand their environments and to communicate and share experiences **//correct?//**. Each sense provides only partial information about the world; we combine information from our different senses with prior knowledge to understand the world. Even vision, the most powerful sense—does not give us enough information to understand the world.

Multimedia computing systems should also combine different sensory media with other knowledge sources to interpret a situation. Multimedia computing is fundamentally about combining information from multiple sources in the context of the problem being solved. This **//characteristic//** distinguishes multimedia from several other disciplines, including computer vision and audio processing, which focus on analyzing one medium to extract as much information as possible.

In multimedia systems, different types of data streams exist simultaneously and the system must process them not as separate streams but as one correlated set of streams that represents information and knowledge useful for solving a problem. The challenge for a multimedia system is to discover correlations in this set of data. It must combine partial information from disparate sources to recover the holistic information in a given context.

# Experience and Information

We experience our physical environment through our natural senses of sight, sound, touch, taste, and smell. From very early **//okay?//**, human children build models of objects and events in the world using all senses available to them. Once these models are in place, we use our senses to function in the physical and social worlds, refining, enhancing, and even developing new models. These models are fundamental to our recognition of objects and events in the world. Modeling an object or event helps us abstract sensory information into a simple symbol. The process of assigning symbols to represent objects and events and then building more complex objects and events using these **//symbols//** is at the heart of *semantics* [8] **//okay to move ref here?//**. Semantics, an area of study in linguistics, relates to the meaning of words, phrases, sentences, and larger units of text.

Although this book does not address semantic theory in detail, we must consider the basic aspects. For our purposes, we consider semantics to be the **study of meaning associated with symbols**. These symbols can be simple atomic symbols or composite symbols built by combining multiple atomic and/or composite symbols. Because our concern is with multimedia signals, these symbols could represent units in different components (such as audio and visual data) or entities as a combination of different media thus resulting in symbols in multimedia.

*Webster*’s dictionary defines experience as the “direct observation of or participation in events as a basis of knowledge.” We experience the world we live in, and we base all of our interactions on those experiences. We learn about the world and accumulate and aggregate our experiences in the form of knowledge. Scientists *experi*ment to test their knowledge and to gain new knowledge. The scientific process relies on experiments to study a hypothesis under different conditions to evaluate its validity. Experiments are fundamental to science and its progress. Humans evaluate their experiments using their sensory processes.

*Communication* is the process of sharing experiences with others. The history of civilization follows the development of our understanding of experiences and how we can share them with other people immediately, even in other parts of the world, as well as with future generations. Many influential innovations and inventions throughout human history are related to communication and the sharing of experiences with people who might be spatially and temporally separated. This process started with the development of languages and has led to the innovations resulting in the World Wide Web.

Information is an “efficient but abstract communication of experience.” We gain knowledge through the bundle of experiences that make up our lives and communicate information about those experiences. Because we don’t communicate the experiences themselves, we lose a vital element of the act of experiencing in the translation of experience to information. Many of us can’t know, for example, what it’s like to hit the game’s winning run, to surf the perfect wave, or to fight a war. We can read about the events, of course, and we can view pictures and videos. But we aren’t present in these technologies, so our experience and hence the knowledge of these events is incomplete.

One of the most important elements in the communication process is developing a dictionary. A dictionary is an exhaustive collection of words in a language. It contains information about the words’ meanings, pronunciations, etymologies, and inflected forms. Thus, a dictionary is a shared and agreed-upon collection of symbols (words) and those symbols’ meanings. Every language is based on a dictionary containing these symbols and rules—the language’s grammar—for using them. Without a dictionary, communication might not happen. Imagine speaking English to a person who is speaking Chinese. Each of you is using a dictionary, but they are two different dictionaries. Communication requires a shared dictionary. In computer science, dictionaries include lists of codes, terms, and keys for use by computer programs.

Many electronic dictionaries now include audio pronunciations. This is the first step toward multimedia dictionaries **//correct?//**. As they evolve, these dictionaries will likely include features in one or more media to form alphabets for that medium and develop approaches to construct multimedia worlds that will represent concepts in the real world.

In multimedia, the basic carrier of meaning (symbols) is not words but units in a particular medium. For example, whereas text uses an alphabet of letters and numbers and speech uses phonemes, audio might include \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ and video might include \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

Consider a simple task involving visual information: Given an image of an object, name the object and list all objects that look like it or are related to it. Try to extend this description to include other common dictionary functions—etymology, pronunciation, examples, and so on. Visual dictionaries are being developed for different applications and for different kinds of visual information, ranging from shapes of curves to complex objects [7]. These dictionaries typically store shapes or objects as well as details about these items in multiple languages. These dictionaries will likely play an progressively more important role in understanding visual information as well as in applying emerging technologies that increasingly use cameras as information devices.

# Perception

Perception is the process of understanding sensory signals or information [6]. Researchers in such areas as psychology and neurophysics have long been analyzing perceptual processes with the goal of understanding them. With the arrival of computing, psychologists and researchers in artificial intelligence **//have been drawn to the study of perception//** in the hopes of developing machines for automatic perception.

Understanding sensory information is an important step in many multimedia systems. Later chapters will discuss important perceptual processes in audio and visual processing. Here, we present some general aspects of perceptual processes.

A perception system aims to **//interpret?//** sensory signals. The system takes sensory signals as input and generates information that is relevant in its application context as output. This process is based on two important sources: signals and relevant knowledge. Figure 1 (Fig:Perception) shows the role of existing knowledge in perception. The system’s output is the combination of the input sensory signal and the knowledge available to the system. Without knowledge, the system cannot produce information, and without the signal, it can only *hallucinate*—that is. **//please briefly explain what you mean by hallucinate in this context.//**. Perception is sometimes considered a controlled hallucination process [3] in which the system starts hallucinating based on the signal and creates multiple hypotheses. It then uses the signal to find information that best supports its hypotheses.

**Existing knowledge**

Data

Information

Figure 1. Perception is the process of recovering meaningful information from sensory data using existing knowledge.



**(a) (b)**

Figure 2: Common illusions: (a) a Dalmatian dog sniffing around and (b) unstable perception. Is it two faces or a vase?

The role of knowledge, in the form of models of objects and events, is not immediately obvious. Some analysis makes it clear, however. We always use the term *recognition* for finding objects in signals, such as images. This term implies that we try to *recognize* objects—that is, we have *models* of objects. Without models, there is no recognition. Models can take many different forms, ranging from concrete and detailed models to abstract models.

The two images in Figure 2 (Fig:illusions) illustrate the importance of models. Figure 2a shows a Dalmatian dog sniffing around. If you don’t know what a Dalmatian dog looks like, you will see only random blobs in this picture; however, if you are familiar Dalmatian dogs, the picture is clear. Figure 2b shows a picture with two interpretations: you can either fit a model of human faces to it and see two faces or see a vase. This demonstrates that your perception system comes up with two hypotheses and admits both as viable, but from a slightly different viewpoint.

# Perceptual Cycle

In all of our activities, we use our senses, brain, and memory to understand our environment, to operate in this environment, to communicate about the environment, and, finally, to build and update our knowledge repositories for efficient and effective use of what we learn. How we use these senses and how we convert this sensory data to information and knowledge has been a source of intrigue for almost all known history.

Here we present some ideas to provide historical context and perspectives on the evolution of this understanding.

Ancient philosophers [5, 13] believed that:

* People derive their understanding of the world indirectly using sensors.
* The fidelity of the model of the world depends on how well a person understands the world.
* People achieve different levels of understanding in terms of their own knowledge structures.
* Nirvana is the highest state of understanding.

Thinkers at this time—approximately 2,000 years ago—clearly recognized that **//people//** must assimilate data from all sensors using **//their//** existing knowledge to form an understanding of the environment. They also recognized that sensors help people understand the world at different levels. We evolve to the highest level of understanding by refining our knowledge structures. Modern philosophers discuss similar ideas and models in the theory of objective reality [10].

To formulate the problem from a computational perspective, we consider Ulric Neisser’s *perceptual cycle* [2,8], introduced in 1976. According to the perceptual cycle, a perceiver builds a model of the world by acquiring specific signals and information to accomplish a task in a natural environment. The perceiver continuously builds and refines a *schema*, which represents the world as the perceiver sees it at a particular instant, based on the signals received up to that point **//correct?//**. The perceiver then decides to get more information to refine the schema for accomplishing its task. This **//action//** sets up the *perceptual cycle*, shown in Figure 3.

**//Applied to computing systems,//** the perceptual cycle involves an agent continuously interacting with the environment using its sensory mechanisms to build a model of the environment that will be useful in its task. At any given instant, the agent has a model of the environment—a schema—based on the data and information it has received so far. The system then decides what more is required to complete the task and how that information could be acquired. Based on this, the agent collects more information using the appropriate sensors.

**Environment**

###### (available information)

**Schema**

**Exploration**

##### Modifies

**Directs**

##### Samples

AGENT

Figure 3. Neisser’s perceptual cycle. The agent (perceiver) receives signals from the environment, which it interprets using the current schema. It modifies the schema based on the results of the analysis and decides whether it needs more information. It continues the cycle until the task is completed.

The perceptual cycle model is conceptually similar to recursive filtering techniques commonly used to estimate a linear dynamic system’s state using observers (sensors) that might provide noisy measurements. These techniques represent the system’s state mathematically as a vector. The state vector represents the values of the parameters used to characterize the linear dynamic system. In system theory [1], these vectors represent the system so that the correct number of control inputs can be applied to the system to bring it into the desired state. In the perceptual cycle, the schema represents the essential parameters required to solve the given task. By comparing the current schema to the final, or desirable, schema, the agent decides its **//next//** action.

However, the perceptual cycle deals with perception that is not **//part of//** a linear dynamic system. This cannot be easily modeled using the current tools of the system theory. Researchers in computer vision have already used some powerful estimators, such as Kalman filters [1], to model aspects of human perception. As technology progresses, more formal tools will likely be developed to represent and construct schema using multimedia data.

# Multimedia Systems

Consider a computing system equipped with multiple sensors working in a physical environment. The system continuously receives information about the environment from these sensors and processes this information in the context of its application. Applications can range from identifying an object to autonomously functioning in a complex dynamic environment. Here, we consider a situation without any specific application. We also consider that for a computing system, the types of data sources are not limited to human sensory modes. Rather, multimedia computing systems could be endowed with sensing capabilities of various types.

Assume we have *S*1, …, *Sn* data streams. These data streams have *K* data types in the form of image sequence, audio stream, motion detector, annotations, symbolic streams, and other relevant and available types. We assume that these streams can be synchronized in both time and space, and thus we can represent them in a common temporal and spatial coordinate system rather than in each sensor’s coordinate system. Further, we assume that metadata *M*1, …, *Mn* for each stream is available from the original sources, which helps us interpret the data stream in the context of the world **//correct?//**. This metadata might include location and type of sensor, viewpoint, angles, or camera calibration parameters. The data stream is rarely directly useful in interpreting the data in relation to the environment. Typically, wemust apply feature detectors to each data stream to obtain features that are relevant in the current environment. Let us represent feature stream *Fij*, where *Fij* is the *j*th feature stream in *Si*.

Given the above data environment, the most fundamental problem that multimedia computing systems must solve is extracting relevant information about the task at hand using data from these disparate sources. This involves many challenging problems, including:

* How do we combine these data streams to obtain the information we need to solve the problem at hand?
* How do we represent this data in the most compact form for communication and storage?
* How do we present this volume of data to users in their computing environments to communicate the intended information?
* What system issues must we solve to deal with the disparate types of data and how the system handles them?

Before we address more specific concepts and techniques related to solving these problems, we discuss some concepts that form the basic fabric of multimedia systems.

## Semantic Gap

In current computer systems, we build abstractions using bits—the most fundamental representation unit of data—and defining concepts that a specific application might need. Building these concepts is easy—we simply define various structures and name and use them as a programmer would **//correct?//**. We are all familiar with concepts such as images, video, and audio signals as they are represented in computers (later chapters discuss these concepts in detail). Human beings—the ultimate users of the computing systems—rarely think in these structures, however. Rather, they usually think in terms of objects and events [11] and build complex concepts based on complex, often ill-defined and uncertain relationships among them. Smeulders et al. define this *semantic gap*—the fundamental gap between the abstractions defined in computing systems and those applied by the users of these systems—as [15]:

“[T]he lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation. A linguistic description is almost always contextual, whereas an image may live by itself.”

Figure 4 shows this semantic gap between the abstractions (images and video) as defined in computers and objects and events as perceived by people.

Bits and Bytes



Alphanumeric Characters

Lists, Arrays, Documents, Images …



**Transformations**

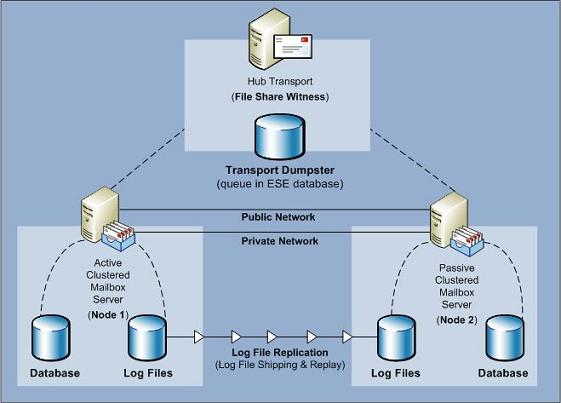
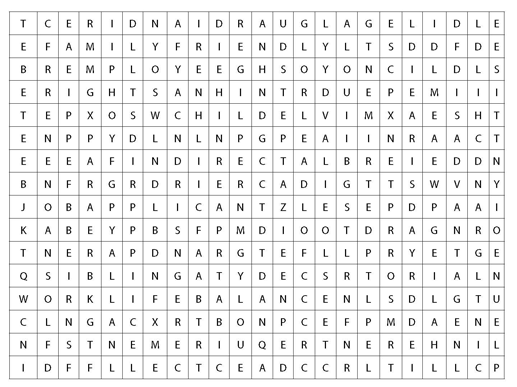
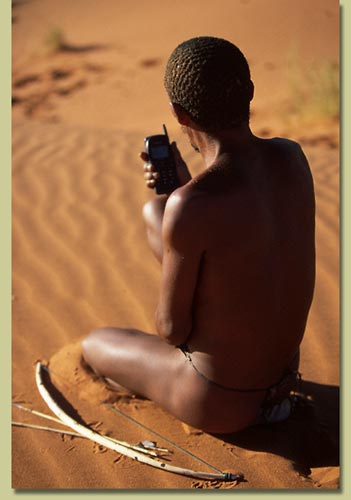


Figure 4: Semantic gap. A large gap exists between how computers represent data such as images in bits and bytes and how people think about images as collection of objects or events.

The main reason for the semantic gap, which can even exist between two persons, is that the physical world contains objects and events. People build models of these objects and events in their minds based on data they receive from their sensors. People learn to abstract this data in terms of objects and events while combining it naturally with data from other sources, such as context and memory. **//Programmers? System designers?//** try to define models of objects and events **//correct?//** in computers using data, which is represented in bits. Thus, these abstractions are based on what can be built in computers using bits. Fundamentally, in computing we define things based on what can be computed, while as humans we learn to abstract what we need to live.

Many concepts and techniques in multimedia computing—from signal analysis approaches in audio processing and computer vision to indexing for multimedia information retrieval—**attempt to bridge** **//correct?//** the semantic gap. In fact, any computing system in which human beings play an integral part must include mechanisms for bridging the semantic gap **//correct?//**. In many mature fields, **//system designers?//** bridge this gap by developing concepts that bring data and humans conceptually close or by developing interfaces that rely on human intelligence to bridge the gap.

Consider search systems. When the task is to search for keywords that can be matched as character strings, a search system can easily give good results. When the task involves searching for something that requires more than string matching, such as some interpretation of data or the user’s intentions, search systems perform poorly. **//Example?//** Most research on improving search engine result relevance tries to bridge the semantic gap by using contextual information to interpret data and detect a user’s intentions.



Figure 5: This photo was taken at Zhujiajiao, Shanghai, China.

## Context and Content

*Content* and *context* are two common terms in multimedia processing and understanding. However, although these terms are used extensively by practitioners and researchers, no rigorous formal definition of them exists. To understand and develop efficient and effective multimedia systems, it is important to understand what they mean and how they are related.

Given a picture, audio, or text file, the content is what is contained in that file. Consider a picture file—say the photo in Figure 5. This photo contains 4,224 × 3,168 color pixels. So, it contains more than 13 million points (pixels), each associated with three color values **//correct?//**. In the most basic form, the file’s content is these 13 million pixel colors in the spatial configuration represented by the pixels’ locations **//correct?//**.

The photo in Figure 5 was taken at a specific time (11:31 a.m. on September 13, **//2009?//**) and location (**Zhujiajiao, Shanghai; latitude** 31 deg 6' 36.34" N**, longitude** 121 deg 2' 59.22" E), using a Nikon Coolpix P6000 camera. No flash was used. The focal length (6.7 mm), ISO (64), and aperture (f/4.7) are also known. The camera captures all of this (and much more) information and stores it in the picture file along with the pixel data using the Exchangeable Image File Format standard (EXIF, http://www.exif.org). This data—the *metadata*—represents the context in which the camera acquired the data in the photo file **//correct?//**.

Context is defined as the interrelated conditions in which some data (the content) is acquired. Almost all digital camera manufacturers use EXIF to store this kind of data.

Some other context parameters might help in understanding the data. For example, in the context of the photo in Figure 5, EXIF can contain the camera model, but having the camera owner’s identity, profile, and calendar information available would also be helpful. In many cases, this information can help identify the objects in the photo and, more importantly, any people in the picture.

Early multimedia research and development was only concerned with data content. Increasingly, researchers are realizing the importance of context. Recent research emphasizes the combination of content and context, taking the view that all information is useful for understanding multimedia [14].

## Metadata

Metadata is, literally, “data about data.” Given some data that represents an audio, photo, or video, its metadata will describe the data’s various characteristics. Some obvious metadata is name, size, creation date, and file type. You can also include any information that might be useful for understanding, storing, transmitting, presenting, or performing any other operation on the file containing the data. Metadata itself is data of some particular type. In this context, it is metadata because it is used to qualify or help some other data.

Metadata has gained popularity with the growth of the Web, **//and in particular, with the use of XML on the Web???//**. XML was designed to be a mechanism to transport and store data. It accomplishes this task by defining a language to describe data. XML tags are data about data. Thus, XML is a language that associates metadata with data explicitly, allowing any program to read it. This facilitates not only transporting and storing data, but also analyzing and understanding it.

The use of metadata in multimedia computing is increasing rapidly. Many approaches based on XML and tags are becoming commonplace, even in audio, images, and video. Techniques to represent metadata will likely evolve in this area. The use of EXIF for stored digital images is a clear example of this trend.

# Objects and Events

When attempting to understand any type of data, an individual tries to determine which aspect of the world the data represents. As discussed previously, perceptual processes depend on prior knowledge about the world we live in. Thus, an important issue is how to represent the world.

**//Around about when?//**,many researchers believed that they could use objects to represent the world. They viewed the world as a collection of objects. Many modern—and not so modern—thinkers have challenged this view [11]. They believe that events play an equally important role. Events represent change in relationships among objects. And these changes are fundamental to understanding the world’s current state. According to emerging views, modeling a dynamic world requires both objects and events. In a sense, objects capture static components of the world, and events complement **these components //???//** by capturing dynamic situations.

Object-oriented approaches have revolutionized many areas of computer science because of the high-level abstractions they offer for design, programming, and even some interactions. Multimedia brings new challenges, however. Multimedia, particularly audio and video, are fundamentally dynamic in nature. They capture signals that represent some attributes of the world as a function of time. In fact, in many applications, even sensors that capture static characteristics of the world, such as temperature at a point in space and time, are used to detect changes in those characteristics as a function of time. A sensor is almost always placed in an environment where some event needs to be detected; the sensor measures some physical attribute that helps in detecting the event.

Many fields in computing have used this concept, but in very different ways. As the use of multimedia increases, a unified approach for event-based thinking will likely evolve [16,17].

Using both objects and events is essential when modeling the real world using powerful computational approaches. Objects and events complement each other: Objects capture attribute and functions of physical objects and related concepts, whereas events represent relationships and changes among those relationships in space and time.

## REFERENCES

1. D. E. Catlin, *Estimation, Control, and the Discrete Kalman Filter*, Springer Verlag, 1988.
2. [Igor A. Chimir](http://findarticles.com/p/search/?qa=Igor%20A.%20Chimir), [Waheeb A. Abu-Dawwas](http://findarticles.com/p/search/?qa=Waheeb%20A.%20Abu-Dawwas), [Mark A. Horney](http://findarticles.com/p/search/?qa=Mark%20A.%20Horney), “Neisser's Cycle of Perception: Formal Representation and Practical Implementation,” *Journal of Computer Science* (Special Issue): 106-111, 2005.
3. M.B. Clowes, “On Seeing Things,” *Artificial Intelligence*, Vol. 2, No. 1, pp 79-112, 1971.
4. ~~http://www.exif.org/~~ **~~//the page stresses that this is an unofficial site; folded into the text. Add as note or further reading instead?//~~**
5. Hermann Kuhn, *The Key To The Center Of The Universe*, Crosswind Publishing, 2001.
6. [James J. Gibson](http://en.wikipedia.org/wiki/J.J._Gibson), *The Ecological Approach to Visual Perception*, Lawrence Erlbaum Associates, 1987.
7. Jonathan Metcalf, *Five Language Visual Dictionary*, 2003.
8. U. Neisser, *Cognition and Reality: Principles and Implications of Cognitive Psychology*, W.H. Freeman, 1976.
9. Jaroslav Peregrin, *Meaning: The Dynamic Turn. Current Research in the Semantics/Pragmatics Interface*, Elsevier, 2003.
10. Karl Popper, “Three World,” the Tanner Lecture on Human Values, delivered by Karl Popper at the University of Michigan, April 7, 1978.
11. A Quinton**,** “[Objects and Events](http://mind.oxfordjournals.org/cgi/reprint/LXXXVIII/1/197),” *Mind*, pp. 197-214, 1979.
12. Irvin Rock, *The Logic of* *Perception*, MIT Press, 1985.
13. L.M. Singavi, *That Which Is: Tattvartha Sutra (Sacred Literature)*, AltaMira Press, 1998.
14. P. Sinha and R. Jain, “Semantics in Digital Photos: A Contenxtual Analysis,” *Int. Journal of Semantic Computing*, Vol. 2, No. 3. pp 311-325, 2008.
15. Arnold Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, “Image Databases at the End of the Early Years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 1, January 2001.
16. A. Scherp and R. Jain, “An Eco System for Semantics,” *IEEE Multimedia*, June 2009.
17. G. Utz Westermann and Ramesh Jain, “Towards a Common Event Model for Multimedia Applications,” *IEEE Multimedia*, January 2007.